动 物 学 研 究 2005, Jun. 26 (3): 225－229 CN 53－1040/Q ISSN 0254－5853
Zoological Research

一种在核苷酸水平检测自然选择的新方法

李 昕[1,2], 陈 宏[1,3], 王 文[2,*]

(1. 西北农林科技大学 动物科技学院, 陕西 杨陵 712100; 2. 中国科学院昆明动物研究所 细胞与分子进化重点实验室, 中德马普青年科学家小组, 云南 昆明 650223; 3. 徐州师范大学 生物技术研究所, 江苏 徐州 221116)

摘要: 非编码区序列在基因表达调控中起着重要作用, 但其在进化过程中是否受到选择作用一直较难检测。最近有一些研究使用平均的核苷酸替换速率与中性序列的核苷酸替换速率的比值 (ω) 作为检测非编码区总体受选择作用的指标; 但是对于非编码区而言, 了解具体哪些核苷酸受到选择作用更具有意义。我们借鉴 Nielsen & Yang (1998) 检测单个氨基酸位点是否受选择作用的思路, 在最大似然法的模型下, 提出一种在核苷酸位点水平上对自然选择作用检测的方法。本方法能够检测在进化过程中对功能分化有重要贡献的核苷酸位点, 包括编码和非编码区。将此方法应用于熟知的受到正选择作用的蛋白编码基因序列 (*HIV-1* 包装蛋白基因编码区), 均能够检测到那些已知的受到正选择的核苷酸 (密码子) 位点, 说明此方法可以有效地在核苷酸位点水平检测选择作用; 又将此方法应用于非编码区 (*CTGF* 基因 5′UTR), 也得到了良好的结果。

关键词: 自然选择; 非编码区; 最大似然法

中图分类号: Q7; Q-332 **文献标识码:** A **文章编号:** 0254－5853 (2005) 03－0225－05

A New Method for Detecting Natural Selection at the Level of Nucleotide Sites

LI Xin[1,2], CHEN Hong[1,3], WANG Wen[2,*]

(1. College of Animal Science and Technology, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi 712100, China; 2. CAS-Max Planck Junior Scientist Group, Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, the Chinese Academy of Sciences, Kunming, Yunnan 650223, China; 3. Institute of Biotechnology, Xuzhou Normal University, Xuzhou, Jiangsu 221116, China)

Abstract: Although noncoding regions play an important role in gene expression and regulation, it is difficult to detect natural selection at this level. Recently, some studies use the ratio (ω) between the nucleotide substitution rate in the detected regions and the nucleotide substitution rate in neutral regions as an indicator to detect natural selection in noncoding regions. However, to the noncoding regions, it's more informative to identify those nucleotide sites under positive selection. We developed a new maximum-likelihood method to detect natural selection at the nucleotide site level and to identify those nucleotide sites that may contribute to functional divergence. This method can be applied to both coding regions and noncoding regions. Appling this method to previous reported genes that subjected to positive selection shows that this method is efficient to detect natural selection on nucleotide sites both in coding regions and noncoding regions.

Key words: Natural selection; Noncoding region; Maximum-likelihood method

在分子水平检测选择作用是近年来进化遗传学中的一个热门话题 (Tajima, 1989; Fu & Li, 1993; Hudson et al, 1987; Mcdonald & Kreitman, 1991; Fay & Wu, 2001; Zhou & Wang, 2004)。对于编码蛋白的基因, 错义替换率 (nonsynonymous substitution rate, Ka) 与同义替换率 (synonymous substitution rate, Ks) 的比值 (Ka/Ks) 被广泛地用来检测基因是否受到选择作用 (Miyata & Yasunaga,

收稿日期: 2004－09－06; 接受日期: 2005－04－01

基金项目: 中德马普青年科学家小组经费资助项目; 中国科学院生物局重要方向性项目 (KSCX2-SW-121); 国家杰出青年科学基金资助项目 (30325016)

* 通讯作者 (Corresponding author), E-mail: wwang@mail.kiz.ac.cn, Tel: 0871－5192979

1980)。在 Ks 为中性替换的假设下，如果 Ka/Ks < 1，说明此基因在进化过程中受到纯化选择（purifying selection）的作用；若 Ka/Ks = 1，即同义替换与错义替换有相同的可能被固定下来，指示此基因经历的是一个中性的进化过程；而如果 Ka/Ks > 1，即某些错义替换具有选择上的优势，其固定的概率和速度要大于同义替换，说明此基因在进化过程中受到了正选择（positive selection)。我们在证明与雄性生殖相关的基因受到强烈正选择作用时即采用了这一思路（Wyckoff et al，2000)。然而由于在进化过程中，一个基因的大部分位点在大部分时间中，一般被认为是受到强的功能限制，因此用整体基因的平均 Ka/Ks > 1 作为受到正选择的标准由于过于严格而缺乏检测效力（Yang & Bielawski，2000；Yang，2002b；Endo et al，1996)。随后人们采用分割法，将整个基因按功能分成几个区域分别计算各个区域的 Ka/Ks，从而提高检验的效力（Hughes & Nei，1988)。然而与功能进化密切相关的位点在 DNA 序列上并不一定是连续排列的，因此这种改进对于检验效力的提高仍然是有限的。最近有研究者提出一系列基于似然法和简约法针对单个氨基酸位点进行选择作用检测的统计学方法，这些方法不但能够更有效地检测基因在进化历程中是否受到选择的作用，并且可以预测出那些在进化过程中对功能分化有重要贡献的、受正选择作用的氨基酸位点（Nielsen & Yang，1998；Fitch et al，1997；Suzuki & Gojobori，1999；Yang & Nielsen，2002)。因此这些方法在很大程度上提高了 Ka/Ks 检验的检测效力，使得对基因适应性进化的检测由整体水平或是区域水平还原到单个氨基酸位点的水平上。

相对于编码蛋白基因分子进化的大量研究，人们对非编码区，包括 RNA 基因、UTRs（untranslated regions)、启动子、内含子和基因间序列的进化研究远远不够。近年来越来越多的研究也表明非编码区同样存在着选择作用（Makalowski et al，1996；Makalowski & Boguski，1998；Larizza et al，2002；Michael et al，2004)。在对非编码区的分析中，类似于前面提到的 Ka/Ks 的标准，被检测区域核苷酸的替换速率（Kd）与中性区域的核苷酸替换速率（Kn）的比值 ω（ω = Kd/Kn）也被用来作为这些区域所受选择作用的指标。$\omega < 1$，$\omega = 1$ 和 $\omega > 1$ 分别作为纯化选择，中性进化和正选择的指示。在我们研究一个年轻的非蛋白编码的 RNA 基因——*sphinx* 基因时，我们使用类似的思路证明了 *sphinx* 基因受到了正选择的作用（Wang et al，2002)。但是，与前面检测编码蛋白基因所存在的问题相似，这种整体替换速率的比较仍然存在检验效力较低的问题，通常只有在选择作用很强时才能检测到。为了解决这个问题，本文尝试将 Nielsen & Yang（1998）的方法从氨基酸位点的水平上推广到单个核苷酸位点的水平上，这样，此方法不仅能够检验编码蛋白基因的选择作用，更重要的是它也能够对非编码区的选择作用进行检测。

1 模型与方法

1.1 位点的分类

考虑多条已排序的核苷酸序列，并且这些序列间的系统发育关系已知。假设序列上各个位点的进化相互独立。我们将所有的核苷酸位点分为 3 类：中性位点、保守位点和正选择位点。r_1、r_2 和 r_3 分别表示待检测序列的中性位点、保守位点和正选择位点的替换速率，r_0 表示中性序列的替换速率，它们的替换速率与中性序列的替换速率的比值定义如下：

$$\begin{cases}\omega_1 = r_1/r_0 = 1\\ \omega_2 = r_2/r_0 = 0\\ \omega_3 = r_3/r_0 > 1\end{cases}$$

我们将保守位点和中性位点的替换速率定为固定值 0 或 1 是为了计算方便，也可以不限制 ω_1 和 ω_2 的值而将其作为未知变量用似然法估计而使模型更符合实际。

1.2 中性替换速率的计算

考虑图 1 所示由 4 条序列构成的系统发育树。我们根据与这 4 条待检验序列相关的中性序列如内含子、假基因、基因间序列，或者同义替换位点，可用 PAML 或 PHYLIP 程序中的最大似然法估计各个枝长（Yang，2002a；Felsenstein，2002)。这样得到 6 条枝的枝长，所得枝长即为各个枝的中性替换速率，分别表示为 V_{15}，V_{25}，...，V_{46}。

1.3 最大似然法估计未知参数

为方便计算，我们仅考虑 Felsenstein 的核苷酸替换模型（Felsenstein，1981)，此方法不考虑转换颠换的差异，而根据序列中某种核酸的比例确定向其突变的概率。其他替换模型计算方法类似，只需添加更多的待估参数即可。由于我们事先并不知道每个位点属于哪一类位点，因此假设整条序列中，

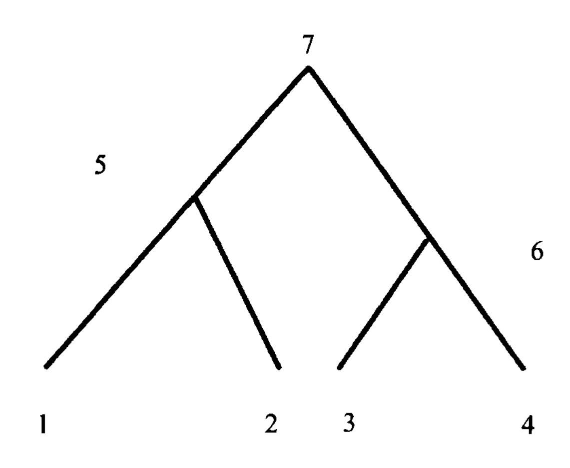


图 1　四条序列构成的系统发育树
Fig. 1　Phylogenetic tree of 4 sequences

中性位点、保守位点和正选择位点所占的比例分别为 p_1，p_2，p_3，$p_1+p_2+p_3=1$。我们采用最大似然法估计 3 个未知参数（ω_3，p_1，p_2）。假设这 4 条序列长度为 n，考虑其中的位点 x_i，根据现在观察到的 4 条序列在此位点上的数据，它的似然值为 $f(x_i)=\sum_{k=1}^{3}p_k\times f(x_i\mid r_k)$，$p_k$ 为此位点属于第 k 类位点的概率，$f(x_i\mid r_k)$ 表示当此位点属于第 k 类位点时，出现现有情况的条件概率，计算它时要考虑每一条枝的进化情况。对于图 1 所示情况来说 $f(x_i\mid r_k)=P_{15}P_{25}P_{57}P_{67}P_{36}P_{46}$，$P_{ij}$ 表示某位点上由结点 i 上的核苷酸替换为结点 j 上的核苷酸的概率。根据 Felsenstein（1981）的核苷酸替换模型：

$$P_{ij}=\begin{cases}\pi_j+(1-\pi_j)\times\exp(-\omega_k V_{ij}/(1-b)),\text{当结点 } i \text{ 与结点 } j \text{ 上的核苷酸相同时；}\\ \pi_j-\pi_j\times\exp(-\omega_k V_{ij}/(1-b)),\text{当结点 } i \text{ 与结点 } j \text{ 上的核苷酸不同时。}\end{cases}$$

其中 π_j 表示第 j 类核苷酸在 4 条序列中的平均频率 j = G，T，C 或者 A；$b=\sum_{j=1}^{4}\pi_j$。这样便计算出了第 i 个位点的似然值。同理，对 n 个位点都做此计算，求它们的对数和，记为 L，$L=\sum_{i=1}^{n}\ln(f(x_i))$，$L$ 即为整条序列总的似然值。调整 ω_3，p_1，p_2 各参数值使 L 最大，估计出 ω_3，p_1，p_2。

1.4　似然比检验

以上是考虑序列中含有正选择位点的情况。同理我们也考虑假设序列中不含正选择位点的情况，即序列仅由中性位点和保守位点组成，仍然可以计算出一个似然值，记为 L_1，这是在中性模型下计算的似然值。然后比较 L 与 L_1 是否有显著差别，即检验用选择模型是否能比中性模型更好地解释数据。由于选择模型包含了中性模型，且比其多 2 个未知参数，我们采用似然比检验，即 $2(L-L_1)$ 近似服从自由度为 2 的 χ^2 分布。

1.5　后验概率的计算

当上述的检验显著后，我们可以计算每个位点属于某一类位点的后验概率，即 $P_{post}=p_k\times f(x_i\mid r_k)$，$P_{post}$ 表示第 i 个位点属于第 k 类位点的后验概率。若某位点属于第三类位点的概率很大，假如大于 0.95 或 0.99，我们就认为此位点在进化过程中受到正选择。

2　结果与分析

2.1　对蛋白编码基因的分析

基于以上方法，我们首先对 *HIV-1* 的包装蛋白基因做分析，由于没有内含子序列，我们用同义替换位点作为中性序列来估计中性替换速率。*HIV-1* 包装蛋白基因序列数据取自 Nielsen & Yang (1998)，计算结果见表 1。

表 1　*HIV-1* 包装蛋白基因分析结果
Tab. 1　Testing result on the *HIV-1* envelope gene

p_1	p_2	ω_3	L	L_1
0.2	0.48	10	−546.762 2	−567.568 1

用 L 和 L_1 进行似然比检验，即检验是否选择模型能更好地符合已知数据，检验结果极显著（$P<0.01$），说明此基因中存在一定比例受到正选择的位点。

对每个位点计算其属于正选择位点的后验概率，结果显示有 4 个位点的后验概率值极显著（表 2）。同时用 Nielsen & Yang（1998）的方法对序列进行分析，亦检测出 4 个极显著位点，其中 3 个位点完全相同，1 个位置上比较接近。说明我们的方法能有效地检测受到正选择的核苷酸位点（表 2）。

2.2　对非编码序列的分析

由于本方法能够直接检测核苷酸位点，因而可以对非编码区进行分析。我们以 *CTGF* 基因（connective tissue growth factor）为例。此前，Larizza et al (2002) 用人、鼠和牛中的 *CTGF* 基因的 mRNA 序列，基于对整个序列的核苷酸替换率与同义替换率的比较，认为此基因的 5′UTR 在进化过程中受到正选择的作用。我们用本方法也对此基因的5′UTR

表 2 *HIV-1* 包装蛋白中预测的正选择位点

Tab. 2 Predicted positively selected sites in the *HIV-1* envelope gene

本文方法 Method of this study			Yang 的方法 Method of Yang*	
位点所在的核苷酸位置 Position of predicted sites in the nucleotide sequence	位点对应的氨基酸的位置 Position of predicted sites in the protein sequence	后验概率 Posterior probability	位点所在的氨基酸位置 Position of predicted sites in the protein sequence	后验概率 Posterior probability
83	28	0.991 624 1	28	0.992 6
107	36	0.991 472 2	31	0.982 1
198	66	0.991 110 7	66	0.992 5
203	68	0.991 400 7	68	0.995 5

* See Nielsen & Yang (1998).

表 3 *CTGF* 基因分析结果

Tab. 3 Testing results on the 5′UTR of the *CTGF* gene

p_1	p_2	ω_3	L	L_1
0.1226	0.3724	70.5	−251.0027	−271.6435

表 4 *CTGF* 基因 5′UTR 中预测的受到正选择作用的位点

Tab. 4 Predicted positively selected sites in the 5′ UTR of *CTGF*

位点所在的核苷酸位置（距翻译起始位点上游的碱基数） Position of predicted sites in the nucleotide sequence (Nucleotide numbers from upstream of start coden)	后验概率 Posterior probability
4	0.973 694 9
12	0.965 945 4
14	0.983 540 1
34	0.955 975 7
38	0.951 803 4
40	0.991 830 6
55	0.974 840 9
73	0.983 540 1
74	0.976 646 7
77	0.956 469 1
90	0.968 406
112	0.965 945 4
114	0.968 406
115	0.975 347 3
117	0.976 646 7

做了分析，各参数的估计值见表 3，选择模型与中性模型的似然比检验极显著（$P<0.01$），说明此基因的 5′UTR 区的确存在正选择作用，并检测出一些后验概率显著的位点（表 4）。更多的近缘物种间该基因序列的比较或人工突变研究可以进一步证实这些结论。

3 讨 论

3.1 对蛋白编码基因选择作用分析的有效性

由对 *HIV-1* 包装蛋白基因 的分析可以看出，本方法所得结果与以氨基酸为分析单位的 Nielsen & Yang (1998) 的方法所得结果基本一致，说明对于编码蛋白基因，本方法是有效的。事实上，如果以同义替换速率作为中性替换速率，那么在错义替换率大于同义替换率的氨基酸位点（密码子），其平均核苷酸替换速率肯定大于中性替换速率。因此，这两种方法在本质上是相似的，应当得到基本一致的结果。

然而，由于密码子使用偏好等原因，同义替换速率未必是中性的（Akashi，1994）。因此，Ka/Ks 大于 1 既可能是由于正选择所驱动的快速的错义替换率，也可能是由于同义替换位点受到选择限制所导致。所以用 Ka/Ks 是否大于 1 作为是否受到正选择的标准可能会引起错误的判断。我们在分析中采用内含子、假基因或基因间序列作为中性序列时可以尽量避免这种错判。尽管寻找一条绝对的中性序列是很困难甚至是不可能的，即便是内含子等人们一般认为的中性序列也可能受到选择作用（Parsch，2003），但是各种类型的序列受到选择作用影响的程度是不同的，因此尽可能准确地估计中性替换速率是本方法得到较可信结果的关键。我们认为，使用待检序列邻近的中性序列应是较好的方法。

3.2 对非编码区选择作用分析的有效性

相对于 Nielsen & Yang (1998) 的方法，本方法的最大优点就是可以对非编码区进行分析，对 *CTGF* 基因 5′UTR 的分析结果说明了其有效性。由于非编码区被认为对基因的表达调控有重要作用，因此该区域所受的选择作用有时对基因的进化是至关重要的。然而如何确定非编码区中的功能序列历来是研究的一个难题，通过本方法预测的受到正选择的位点可以为进一步的功能研究提供一个理论参考。

3.3 存在的问题

由于核苷酸的种类只有4种，因此对于 n 条已对齐（alignment）的序列来说，每个核苷酸位点理论上可以有 4^n 种可能。而组成蛋白质的氨基酸有20种，编码20种氨基酸的密码子有61个，因此，以氨基酸（密码子）为分析单位时，会利用3个核苷酸位点的信息，其每个氨基酸位点理论上有 61^n 种可能。这样对于编码蛋白基因来说，这种基于核苷酸水平的分析方法与以氨基酸为分析单位的方法相比，对于区分不同位点间的选择作用分辨率相对不高。不同位点的后验概率值有时会非常接近。另外，对于一些核苷酸替换较快的序列，常会使预测正选择位点失效，使得每个发生替换位点的后验概率计算都为1，从而难以确定具体受正选择作用的位点。即便如此，这时选择模型与中性模型的似然比检验仍然是有效的，仍然可以探讨被检测序列整体上受选择作用的情况。总体而言，本方法在有大量近缘序列时会比较有效，在序列较少或序列间分歧（divergence）很大时，要慎重使用。

参考文献：

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy [J]. *Genetics*, **136**: 927-935.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate [J]. *Mol. Biol. Evol.*, **13**: 685-690.

Fay JC, Wu CI. 2001. The neutral theory in the genomic era [J]. *Curr. Opin. Genet. Dev.*, **11**: 642-646.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach [J]. *J. Mol. Evol*, **17**: 368-376.

Felsenstein J. 2002. PHYLIP (Phylogeny Inference Package), version 3.6 [CP]. Department of Genome Sciences, University of Washington, Seattle.

Fitch WM, Bush, RM, Bender CA, Cox NJ. 1997. Long term trends in the evolution of H (3) HA1 human influenza type A [J]. *Proc. Natl. Acad. Sci.*, **94**: 7712-7718.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations [J]. *Genetics*, **133**: 693-709.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data [J]. *Genetics*, **116**: 153-159.

Hughes AL, Nei M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class Ⅰ loci reveals overdominant selection [J]. *Nature*, **335**: 167-170.

Larizza A, Makalowski W, Pesole G, Saccone C. 2002. Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyls and rodent gene pairs [J]. *Comput. Chem.*, **26**: 479-490.

Makalowski W, Boguski MS. 1998. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2820 orthologous rodent and human sequences [J]. *Proc. Natl. Acad. Sci.*, **95**: 9407-9412.

Makalowski W, Zhang J, Boguski MS. 1996. Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences [J]. *Genome Res.*, **6**: 846-857.

Mcdonald J, Kreitman M. 1991. Adaptive protein evolution at *adh* locus in *Drosophila* [J]. *Nature*, **351**: 652-654.

Michael H, Fang S, Wu CI. 2004. Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes [J]. *Mol. Biol. Evol.*, **21** (2): 374-383.

Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: A method for estimating evolutionary rate of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications [J]. *J. Mol. Evol.*, **16**: 23-36.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positive selected amino acid sites and applications to the *HIV-1* envelope gene [J]. *Genetics*, **148**: 926-936.

Parsch J. 2003. Selective constraints on intron evolution in *Drosophila* [J]. *Genetics*, **165**: 1843-1851.

Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites [J]. *Mol. Biol. Evol.*, **16**: 1315-1328.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism [J]. *Genetics*, **123**: 585-595.

Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster* [J]. *Proc. Natl. Acad. Sci.*, **99**: 4448-4453.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man [J]. *Nature*, **20**: 304-309.

Yang Z. 2002a. Phylogenetic Analysis by Maximum Likelihood (PAML), version 3. 13d [CP]. Department of Biology, University College London, London.

Yang Z. 2002b. Inference of selection from multiple species alignments [J]. *Curr. Opin. Genet. Dev.*, **12**: 688-694.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation [J]. *Trends. Ecol. Evol.*, **15** (12): 496-503.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages [J]. *Mol. Biol. Evol.*, **19**: 908-917.

Zhou Q, Wang W. 2004. Detecting natural selection at the DNA level [J]. *Zool. Res.*, **25** (1): 73-80. [周　琦，王　文. 2004. DNA水平自然选择的检测. 动物学研究，**25** (1): 73-80.]